**Enhancing Heart Disease Prediction Using ADA Boost Algorithm and XG Boost Algorithm**

**Authors:** Archita Gupta, Sonali Raghuwanshi, Abhinav Shrivastava, Tina Chelwani and Priyanshi under the guidance of **Dr. Manikandan**

**Abstract**

Cardiovascular disease is the group of diseases which affects our heart and blood vessels. These diseases have multifactorial origins, including genetic predispositions, poor dietary habits, physical inactivity, and the presence of comorbidities like hypertension and diabetes. Despite significant advancements in medical research and healthcare, heart disease remains prevalent due to factors such as longer life expectancy, urbanization, poor-quality diets, and sedentary lifestyles.

Furthermore, disparities in healthcare access and preventive measures exacerbate the situation, particularly among socioeconomically disadvantaged populations. In response, there has been a growing focus on leveraging advanced technologies such as machine learning, big data analytics, and predictive modelling to address these challenges. These methodologies utilize computational tools and large datasets to improve risk assessments, optimize treatment plans, and reduce the burden of cardiovascular diseases on both individuals and healthcare systems.

This research focuses on the application of robust data preprocessing techniques, including handling class imbalance, finding missing values, normalization using IQR method and outliers removal. Furthermore, feature engineering techniques are utilized to introduce new relevant features, such as cholesterol, systolic blood pressure and diastolic blood pressure to improve model accuracy.

Additionally, the study integrates the Ada Boost Classifier and XG Boost Classifier techniques for modelling, aiming to improve the accuracy of heart disease prediction. The models were evaluated based on the accuracy, precision, recall, and F1-score. The XG Boost model performed best with 72.82% of accuracy.

**Keywords:** IQR, ADA Boost, XG Boost

1. **Introduction**

Heart disease is one of the leading causes of health problems worldwide. It includes various conditions like heart failure, arrhythmias, and coronary artery disease. These diseases are often caused by genetics, lifestyle habits, and conditions like diabetes and high blood pressure. Despite advances in medical research, the number of heart disease cases remains high. Factors such as aging, urbanization, unhealthy diets, and lack of physical activity contribute to this rise. The burden is greater in underprivileged communities with limited access to healthcare. By using big data and computational tools, we can identify risks, improve treatment options, and reduce the impact of heart disease on people and healthcare systems. Machine learning plays a crucial role in this challenge by analysing data to uncover patterns and predict risks. It helps identify people at risk, supports early diagnosis, suggests personalized treatments, and monitors heart health. By integrating machine learning with healthcare, we can make better decisions and improve patient outcomes, ultimately reducing the global impact of heart disease.

1. **Literature review**

In [2017](https://www.researchgate.net/publication/318604547_Heart_Disease_Prediction_System_Using_Random_Forest)Yeshvendra Singh, Nikhil Sinha, and SanjayKumar Singh trained the Random Forest model using on the dataset from the Cleveland heart disease study includes 303 instances, each with 14 attributes (13 features and 1 target variable). related to cardiovascular health. By utilizing well-set parameters, the Random Forest achieved an accuracy of 85.81%. The model's performance was validated using 10-fold cross-validation, demonstrating the machine learning effectiveness in improving cardiovascular disease prediction. The given approach offers significant potential for saving lives through early diagnosis. In [[2019]](https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072/pdf), Abhijeet Jagtap, Priya Malewadkar developed a heart disease prediction system using machine learning techniques, including Decision Trees, Naive Bayes, and Neural Networks. Their research emphasized the need for mining healthcare data to uncover hidden patterns, improving medical decision-making. They also developed a smartphone-based application for predicting Ischemic Heart Disease risk using clinical data from 787 patients, showcasing practical machine learning applications in healthcare. Baban Uttamrao Rindhe and colleagues [[2021]](https://www.researchgate.net/publication/351545128_Heart_Disease_Prediction_Using_Machine_Learning) applied machine learning techniques like Artificial Neural Networks (ANN), Random Forest, and Support Vector Machines (SVM) to predict heart diseases, which are a leading cause of death globally. These algorithms help analyze large medical datasets, improving prediction accuracy and assisting healthcare professionals in early diagnosis. By identifying risk factors such as lifestyle and stress, the research aims to reduce healthcare costs and enhance patient outcomes. In [[2021]](https://www.researchgate.net/publication/352200597_Heart_disease_prediction_using_data_mining), authors applied data mining techniques to predict heart disease using the Cleveland dataset with 14 attributes. They implemented Naïve Bayes and Decision Tree classifiers to analyze and identify patterns related to heart disease. Key parameters included smoking, cholesterol levels, family history, and physical activity. The dataset consisted of 303 records, and the authors compared the effectiveness of these algorithms for heart disease prediction.In [[2023]](https://www.mdpi.com/1999-4893/16/2/88), authors focused on enhancing model accuracy by employing k-modes clustering for preprocessing. They utilized a Kaggle dataset of 70,000 instances and applied GridSearchCV to fine-tune the parameters, achieving high accuracy with techniques such as Decision Tree, Random Forest, XGBoost and Multilayer Perceptron. The best-performing model, Multilayer Perceptron, achieved an accuracy of 87.28% with cross-validation. The study emphasizes the potential of machine learning to improve heart disease prediction, particularly in reducing fatalities and healthcare costs globally. The study by Ch. Kiran Babu and colleagues [[2024](https://jnao-nu.com/Vol.%2015,%20Issue.%2001,%20January-June%20:%202024/58_online.pdf)] focuses on heart disease prediction using machine learning. They proposed an innovative Principal Component Heart Failure (PCHF) feature engineering method that enhance prediction accuracy. Nine machine learning algorithms, including decision trees, logistic regression, and XGBoost, were used, with cross-validation validating the results. The decision tree model outperformed others, demonstrating the effectiveness of optimized feature selection and hyperparameter tuning in heart disease detection. The study by Haikun Guo [[2024]](https://www.scitepress.org/Papers/2023/128007/128007.pdf) evaluates the efficacy of five machine learning models—Logistic Regression, K-Nearest Neighbor, Random Forest, Decision Tree, and Gradient Boosting—on predicting Coronary Heart Disease (CHD). The study addresses The drawbacks of traditional statistical techniques, highlighting the application of machine learning in managing multidimensional data.. After training and testing on the Framingham Heart Study dataset, the models were assessed using metrics like accuracy, specificity, and sensitivity. The results showed that Random Forest and Gradient Boosting were particularly effective in predicting CHD, offering valuable insights for early diagnosis and preventive healthcare strategies.

1. **Proposed Methodology**

The proposed method aims to enhance the accuracy and performance of machine learning models for heart disease prediction by leveraging data preprocessing and advanced algorithms like AdaBoost and XG Boost. The process begins with data cleaning, addressing inconsistent values, missing data, and duplicates to ensure a high-quality dataset. Outliers are managed using the IQR method, improving data reliability and ensuring accurate predictions. Feature importance is identified through model-based analysis to focus on the most impactful features. Ensemble learning methods, AdaBoost and XG Boost, are utilized to capture complex relationships among features, boosting prediction accuracy and robustness. This approach delivers efficient and reliable predictions, supporting improved healthcare decisions and patient outcomes.

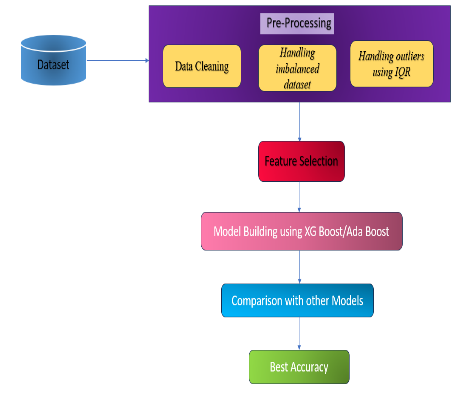


Fig. 1. Heart Disease Prediction Workflow

* 1. ***Data Collection***

The dependent variable in the dataset indicates whether a person has heart disease, based on 13 key predictor variables. These features include medical data such as systolic and diastolic blood pressure, cholesterol levels, glucose levels, and activity levels, along with demographic and behavioural factors like gender, smoking habits, and alcohol consumption. The dataset comprises 10,000 records, providing a robust foundation for training machine learning algorithms to predict heart disease. By analysing these variables, the model uncovers critical insights into the relationship between heart disease and various health-related or lifestyle factors.

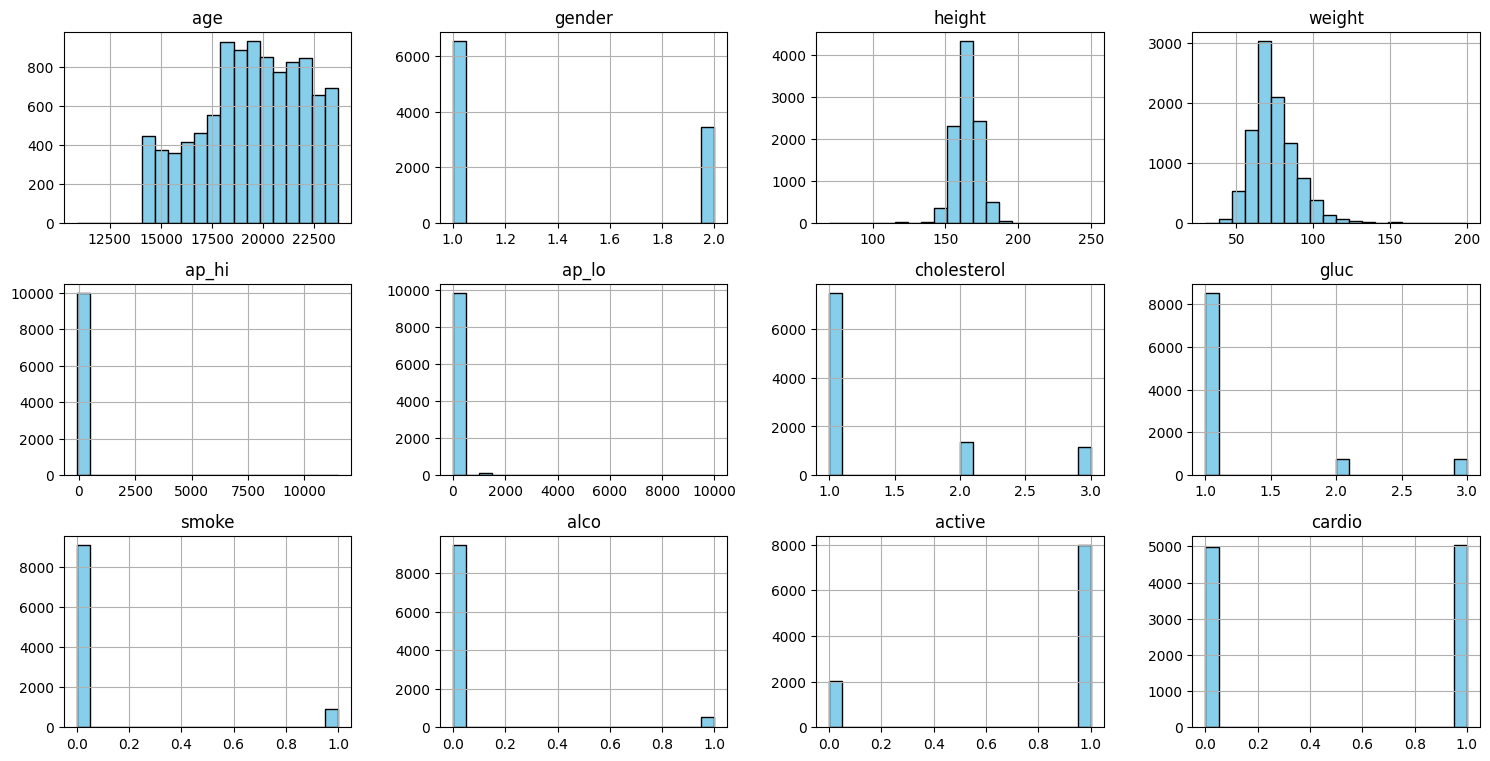


Fig. 2. Histograms of numeric columns in heart disease dataset

* 1. ***Data Preprocessing***
     1. ***Data cleaning***

The dataset used in this project had no missing values, ensuring its completeness and eliminating the need for techniques like imputation or deletion of rows/columns. This simplifies the data preprocessing steps and ensures the dataset's integrity for analysis. Missing values can often create challenges in data analysis and machine learning, leading to biased or inaccurate results. However, the absence of such issues in this dataset allows for a more direct and reliable application of machine learning techniques.

* + 1. ***Handling imbalanced dataset***

In this research, the dataset was inherently balanced with 5,030 cases classified as negative and 4,969 as positive, eliminating the need for oversampling or undersampling techniques like SMOTE (Synthetic Minority Over-sampling Technique). A balanced dataset ensures that the machine learning model can learn effectively from all classes without being biased towards the majority class.

* + 1. ***Handling Outliers using IQR***

Outlier identification and removal are crucial steps in enhancing the reliability and efficiency of machine learning models. Outliers, which are extreme values that significantly deviate from the rest of the data, can distort the analysis and negatively affect model performance. The Interquartile Range (IQR) method is commonly used to detect outliers. In this approach, any data point greater than Q3 + 1.5 \* IQR or less than Q1 − 1.5 \* IQR is considered an outlier, where Q1 and Q3 represent the first and third quartiles, respectively, while IQR denotes the interquartile range. In our work, outliers were specifically identified and removed from the parameters age, height, and weight. By eliminating these extreme values, we ensured that the dataset better represents the majority of the observations, which in turn reduces bias and improves the accuracy of the machine learning model. The use of the IQR method in this context allowed us to create a cleaner and more reliable dataset.

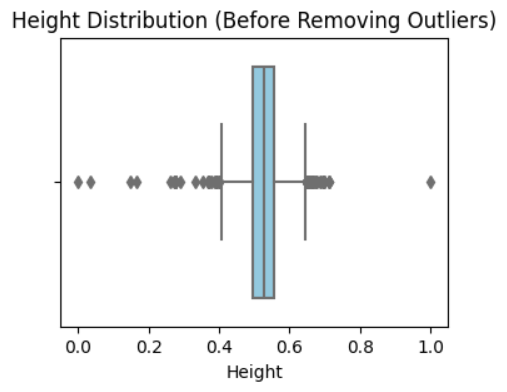
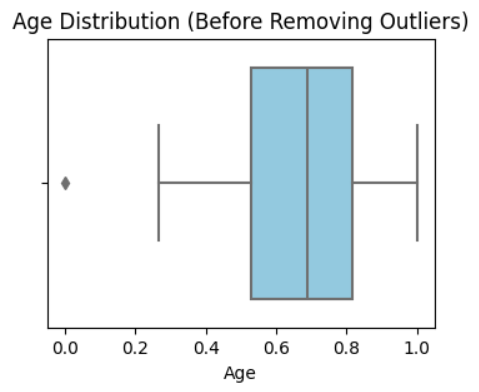
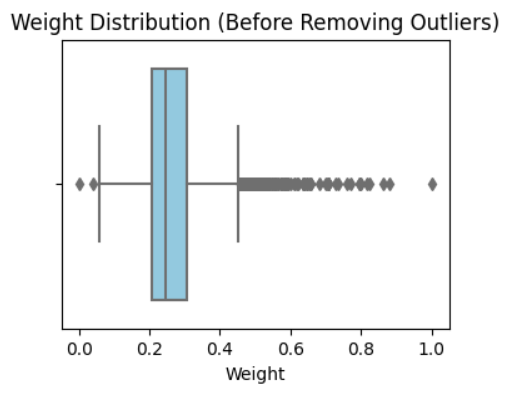
 

Fig3: Before Handling Outliers

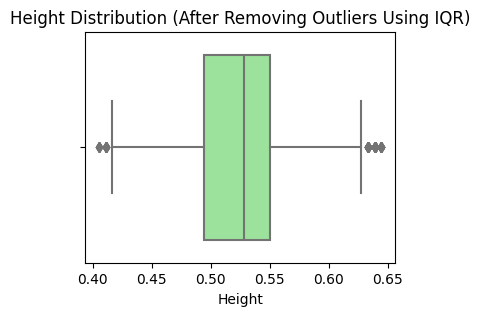
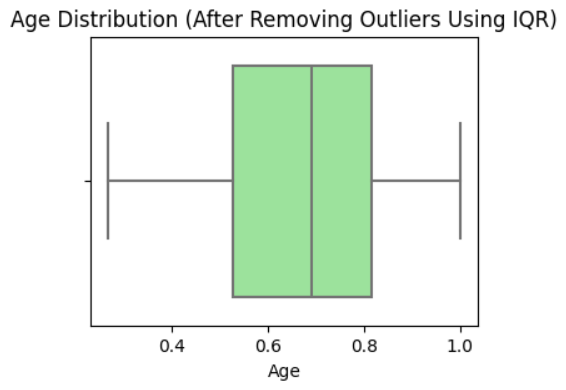
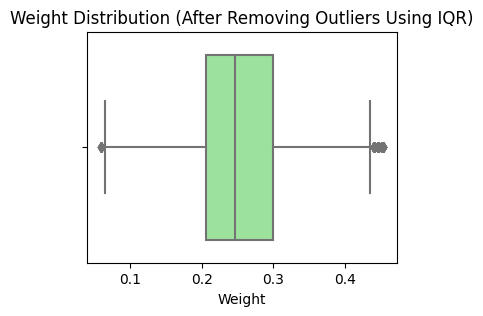
 

Fig4: After Handling Outliers

* 1. ***Feature Selection***

In our feature selection process, we focused on selecting the most relevant features to enhance the performance of the heart disease prediction model. We carefully evaluated features such as gender, systolic blood pressure, diastolic blood pressure, cholesterol levels, smoking status, alcohol consumption, and physical activity for their importance in predicting heart disease. By eliminating irrelevant or redundant features, we identified the five most important features: systolic blood pressure, diastolic blood pressure, cholesterol levels, physical activity, and gender. This reduced the dimensionality of the data, minimizing overfitting and improving the model’s generalization ability.

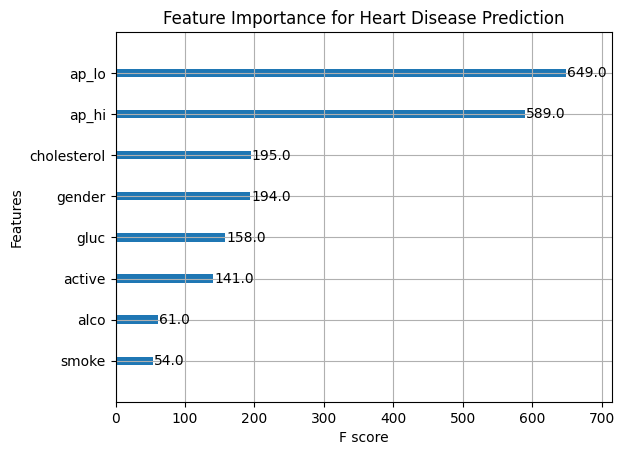


Fig 5: Important Features in XG Boost Algorithm

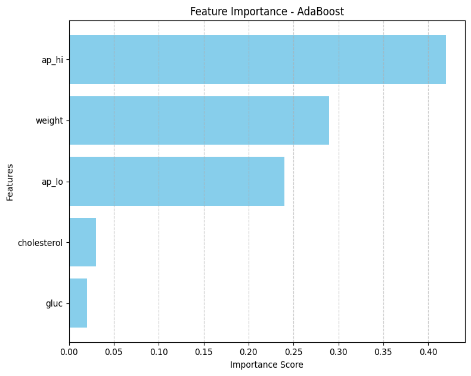


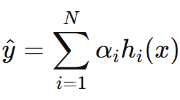
Fig 6: Important Features in Ada Boost Algorithm

* 1. ***Model Building***

In our model-building process, we employed AdaBoost and XG Boost, two powerful ensemble learning methods, to predict heart disease. These models combine the predictions of multiple weak learners (decision trees) to create a robust predictive model. Both AdaBoost and XG Boost improve prediction accuracy by iteratively adjusting the weights of misclassified data points, focusing more on difficult cases in subsequent iterations.

* + 1. ***AdaBoost***

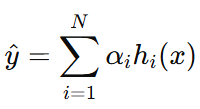
The AdaBoost algorithm creates a strong predictive model by combining multiple weak classifiers. It works by sequentially applying weak learners (e.g., decision trees) to the dataset. During each iteration, the model adjusts the weights of incorrectly classified instances, focusing more on them in the next iteration. AdaBoost’s prediction is given by:



where ŷ​ is the predicted value, N is the number of weak classifiers (trees), hi(x) is the prediction from the i-th weak classifier, and αi ​ is the weight of the classifier based on its performance.

* + 1. ***XG Boost***

XG Boost is an advanced version of gradient boosting that applies gradient boosting trees with optimization techniques. It builds trees sequentially, with each tree attempting to correct the errors made by previous ones. By using regularization techniques and handling missing data efficiently, XG Boost provides faster training and better predictive performance. The prediction from an XG Boost model is:



where ŷ​ is the predicted value, N is the number of trees, hi(x) is the prediction of the i-th tree, and αi ​ is the weight assigned to each tree.

1. **Results and Discussion**

***4.1 Performance Assessments***

* + 1. ***Feature Selection***

The feature selection process utilizes the distinct capabilities of XG Boost and AdaBoost to identify crucial predictors. Both methods agree on significant features like systolic blood pressure, diastolic blood pressure, cholesterol level, and glucose level, highlighting their consistent importance across different algorithms. AdaBoost uniquely identifies body weight as a critical feature, leveraging its iterative focus on improving classification accuracy. On the other hand, XG Boost captures physical activity, emphasizing its ability to identify interactions in the dataset. By combining insights from both models, the approach ensures a well-rounded feature selection strategy that balances interpretability, precision, and the ability to generalize effectively. The model’s performance, with evaluation metrics like accuracy or ROC-AUC, reflects good prediction accuracy given the complexity of the dataset. This demonstrates how well the feature selection methods, using AdaBoost and XG Boost, identify a small but impactful set of features. These selected features contribute to the model’s performance while also enhancing its interpretability. Further research, exploring different validation techniques and feature selection criteria, would help confirm the generalizability and robustness of these features in practical applications.

* + 1. ***Model Building using XG Boost and Ada Boost***
       1. ***Ada Boost***

Ada boost (Adaptive Boosting) is another powerful ensemble learning algorithm that enhances the performance of weak classifiers by combining them to create a strong classifier. It works by focusing on the misclassified samples from previous iterations and giving them more weight in subsequent iterations. This adaptive strategy allows Ada boost to correct errors from earlier classifiers and improve the overall model accuracy. Ada boost is particularly effective in reducing bias, making it a suitable choice for handling complex datasets with imbalances or noise. It helps to improve predictive accuracy by adjusting the weights iteratively, ultimately resulting in a robust and reliable mode.

Table 1 : Model Performance Metrics

|  |  |
| --- | --- |
| **Performance Metrics** | |
| **Metrics** | **Values** |
| Accuracy | 0.7286 |
| Precision | 0.77 |
| Recall | 0.64 |
| F1 score | 0.70 |

When comparing AdaBoost to other algorithms, it performs well in key metrics like Precision, F1 Score, Recall and Accuracy. For example, AdaBoost achieved 73% Accuracy, indicating the ratio of instances correctly classified. The model’s capability to accurately detect heart disease cases is shown by its 77% Precision, and the 64% Recall indicates how well it detects actual positive cases. The F1 score of 0.70 balances the precision and recall, indicating good overall performance in heart disease prediction.

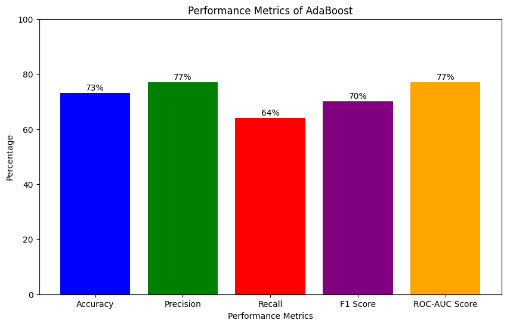


Fig.7 :. A bar graph for performance metrics

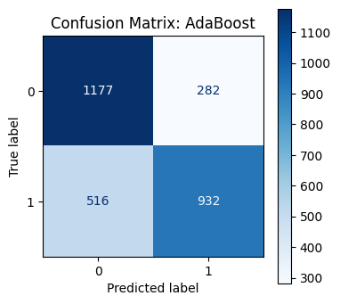


Fig 8: Confusion Matrix for model Building

* + - 1. ***XG Boost***

XGBoost (Extreme Gradient Boosting) is a powerful ensemble learning algorithm commonly used for classification and regression tasks in machine learning. Itworks by building a series of decision trees sequentially, where each tree attempts to correct the errors made by the previous one. XGBoost uses a boosting approach, where each tree's contribution is weighted based on its performance. This iterative process enhances predictive accuracy and minimizes errors. To combat overfitting, XGBoost employs techniques like regularization, which controls the complexity of the model, and early stopping, which halts training when performance on a validation set starts to degrade. These strategies make XGBoost highly effective, especially for high-dimensional datasets with complex feature interactions, ensuring robustness and high performance.

Table 2: Model Performance Metrics

|  |  |
| --- | --- |
| **Performance Metrics** | |
| **Metrics** | **Values** |
| Accuracy | 0.7282 |
| Precision | 0.79 |
| Recall | 0.62 |
| F1 score | 0.69 |

When comparing to Ada Boost and other algorithms, XGBoost typically demonstrates strong performance across key metrics such as Accuracy, Precision, Recall, F1 Score, and ROC-AUC Score. For example, an XGBoost model in this case achieved an accuracy of 0.7282, which represents the proportion of accurately identified samples. Model's proficiency in correctly identifying true affirmative cases is shown by its 79% Precision, while the 62% Recall reveals how effectively the model captures actual positive cases in the predictions. Additionally, the 77% ROC-AUC score highlights the model’s ability to distinguish between the two classes, further demonstrating its strong predictive power in heart disease detection.

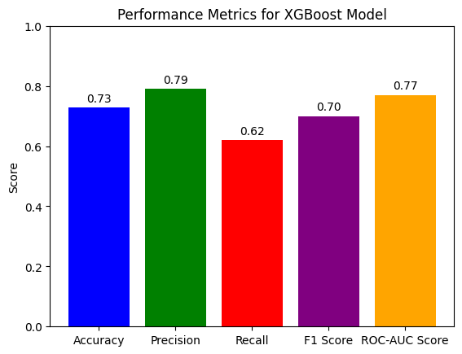


Fig.9 : A bar graph for performance metrics for XG Boost

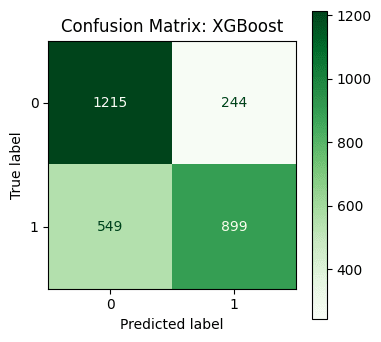


Fig 10: Confusion Matrix for model Building

* 1. ***Comparison of Proposed Method and Other methods on Heart Disease Prediction***

When comparing the research works published in recent years on heart disease prediction, several important contributions have been made. In 2021, Baban Uttamrao Rindhe, Nikita Ahire, Rupali Patil, and Shweta Gagare proposed a heart disease prediction model using machine learning algorithms, including Support Vector Machine (SVM), Artificial Neural Networks (ANN), and Random Forest Classifier. After preprocessing the data, their models achieved the following accuracy: SVM at 84.0%, ANN at 83.5%, and Random Forest at 80.0%. Similarly, R. Fadnavis, K. Dhore, D. Gupta, J. Waghmare, and D. Kosankar, in the same year, focused on data mining for heart disease prediction, employing Naive Bayes and Decision Trees, which achieved accuracies of 85.25% and 81.97%, respectively. In 2020, Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, and Preeti Nagrath created a cardiovascular disease detection model using Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest Classifier. They achieved an accuracy of 87.5%, with KNN performing at 88.52%. In 2019, Abhijeet Jagtap, Priya Malewadkar, Omkar Baswat, and Harshali Rambade employed SVM, Logistic Regression, and Naïve Bayes, achieving accuracies of 64.4%, 61.45%, and 60%, respectively. Their model was selected for a web-based heart disease prediction application based on the performance of the SVM algorithm. In our study, we implemented two powerful algorithms, AdaBoost and XGBoost, for heart disease prediction. Through meticulous data preprocessing and preparation, including outlier detection using the Interquartile Range (IQR) method and feature scaling, we achieved reliable results. AdaBoost performed with an accuracy of 72.86%, recall of 64%, and an F1 score of 70%, while XGBoost achieved similar results with an accuracy of 72.82%, recall of 62%, and an F1 score of 69%. These findings show the effectiveness of ensemble models like AdaBoost and XGBoost in heart disease prediction tasks, aligning with trends observed in recent studies but contributing new insights through our data preparation and model tuning approaches. When comparing AdaBoost to other algorithms, it performs well in key metrics like Recall, Accuracy, F1 Score and Precision. For example, AdaBoost achieved 73% Accuracy, reflecting the rate of correctly classified record. The capability to precisely detect heart disease cases is shown by its 77% Precision, and the 64% Recall indicates how well it detects actual positive cases. The F1 score of 0.70 balances the precision and recall, indicating good overall performance in heart disease prediction.

Table3: Comparative Performance of other models

|  |  |  |
| --- | --- | --- |
| **Author** | **Method Used** | **Accuracy** |
| Baban Uttamrao et al. (2021) | Random Forest | 80.0%. |
| R. Fadnavis  et al.  (2021) | Naive Bayes and Decision Trees | 81.97% |
| Harshit Jindal et al.(2020) | K-Nearest Neighbors (KNN), and Random Forest Classifier. | 87.5% |
| Abhijeet Jagtap et al.(2019) | SVM, Logistic Regression, and Naïve Bayes | 60% |
| Our Study | XG Boost | 72.82% |

1. **Conclusion**

In conclusion, this work demonstrates how to improve the performance of heart disease prediction models. We ensured that the dataset was clean and ready for analysis by handling missing values and removing outliers using the IQR method. By carefully selecting the most important features, we reduced the complexity of the model and improved its efficiency. We used Random Forest and XG Boost for model building, which enhanced the accuracy of predictions. The steps taken in this study clearly show how each phase contributes to building a more reliable and accurate model. Future work could explore these methods with larger datasets and in different fields to improve model interpretability and performance. This research contributes to the field by providing a framework that makes machine learning models more practical and effective in real-world applications.

1. **References**
2. Yeshvendra Singh, Nikhil Sinha, Sanjay Kumar Singh, "Heart Disease Prediction System Using Random Forest," *Communications in Computer and Information Science*, July 2017, DOI: 10.1007/978-981-10-5427-3\_63, Conference: International Conference on Advances in Computing and Data Sciences.
3. Baban Uttamrao Rindhe, Nikita Ahire, Rupali Patil, Shweta Gagare, "Heart Disease Prediction Using Machine Learning," *International Journal of Advanced Research in Science Communication and Technology*, May 2021, DOI: 10.48175/IJARSCT-1131.
4. R. Fadnavis, K. Dhore, D. Gupta, Jaysinh Waghmare, "Heart disease prediction using data mining," *Journal of Physics Conference Series*, May 2021, DOI: 10.1088/1742-6596/1913/1/012099, License: CC BY 3.0.
5. Chintan M. Bhatt, Parth Patel, Tarang Ghetia, Pier Luigi Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," *Algorithms*, 2023, DOI: 10.3390/a16020088, Published: 6 February 2023.
6. CH. Kiran Babu, M. Iswarya, R. Manikanta Kumar, M. Pavan Sai, "Effective Feature Engineering Technique for Heart Disease Prediction with Machine Learning," *Journal of Nonlinear Analysis and Optimization*, Vol. 15, Issue 1, 2024, ISSN: 1906-9685.
7. Abhijeet Jagtap, Priya Malewadkar, Omkar Baswat, Harshali Rambade, "Heart Disease Prediction using Machine Learning," *International Journal of Research in Engineering, Science and Management*, Volume-2, Issue-2, February 2019, ISSN (Online): 2581-5792.